Finding Intermediary Topics Between People of Opposing Views: A Case Study

Eduardo Graells-Garrido*
Telefónica I+D
Santiago, Chile

Mounia Lalmas
Yahoo Labs
London, UK

Ricardo Baeza-Yates
Yahoo Labs
Sunnyvale, USA

ABSTRACT

In micro-blogging platforms, people can connect with others and have conversations on a wide variety of topics. However, because of homophily and selective exposure, users tend to connect with like-minded people and only read agreeable information. Motivated by this scenario, in this paper we study the diversity of intermediary topics, which are latent topics estimated from user generated content. These topics can be used as features in recommender systems aimed at introducing people of diverse political viewpoints. We conducted a case study on Twitter, considering the debate about a sensitive issue in Chile, where we quantified homophilic behavior in terms of political discussion and then we evaluated the diversity of intermediary topics in terms of political stances of users.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

Keywords

Social Networks; Topic Modeling; Homophily; Political Diversity.

1. INTRODUCTION

Social research has shown that, while everyone indeed has a voice, people tend to listen and connect only to those of similar beliefs in political and ideological issues, a cognitive bias known as homophily [19]. This bias is present in many situations, and it can be beneficial, as communication with culturally alike people is easier to handle. However, the consequences of homophily in ideological issues are prominent, both off- and on-line. On one hand, groups of like-minded users tend to disconnect from other groups, polarizing group views. On the other hand, Web platforms recommend and adapt content based on interaction and network data of users, *i. e.*, who is connected to them and what they have liked before. Because algorithms want to maximize user engagement, they recommend content that reinforces the homophily in behavior and display only agreeable information. Such biased reinforcement, in turn, makes computer systems to recommend even more polarizing content, confining users to *filter bubbles* [20].

One way to improve the current situation is to motivate users to read challenging information, or to motivate a change in behavior through recommender systems. However, this "direct" approach has not been effective as users do not seem to value political diversity or do not feel satisfied with it [1], a result explained by *cognitive dissonance* [14], a state of discomfort that affects persons confronted with conflicting ideas, beliefs, values or emotional reactions. Conversely, Graells-Garrido *et al.* [16] proposed an "indirect" approach, by taking advantage of partial homophily to suggest similar people, where similarity is estimated according to *intermediary topics*. Intermediary topics are defined as non-conflictive shared interests between users, *i. e.*, interests where two persons of opposing views on sensitive issues could communicate and discuss without facing challenging information in a first encounter. According to the primacy effect in impression formation [2], first impressions matter, making such intermediary topics important when introducing people. In recommender systems, recommendations based on intermediary topics would indirectly address the problem of exposing people to others of opposing views in a non-challenging context.

In this work, we extend the definition of intermediary topics [16]. In addition, we formally evaluate this redefinition by considering the following research question: *are intermediary topics more diverse in terms of political stances and language than non-intermediary topics?* We approach this question by performing a case study on the micro-blogging platform Twitter, with users who discussed sensitive issues, *i. e.*, ideological or political themes that would make people reject connecting or interacting with others. In particular we focus on the analysis of discussion around *abortion* in Chile. Chile has one of the strictest abortion laws in the world [24], yet at the same time a majority of population is in favor of its legalization [10], making it a controversial topic suitable for analysis. Our contributions include a quantification of the homophilic structure of discussion around this topic in Chile, and a confirmation of the diversity of people with respect to political stances in intermediary topics.

This paper is organized as follows: after reviewing the background work (§ 2), we define the methods and concepts needed to study intermediary topics (§ 3). Then, we perform a case study in Chile (§ 4). Finally, we discuss results and implications (§ 5).

2. BACKGROUND

Homophily is the tendency to form ties with similar others, where similarity is bound to many factors, from sociodemographic to behavioral and intra-personal ones (see a literature review by McPherson *et al.* [19]). In micro-blogging platforms, homophily has been observed in terms of political leaning [5]. Because of homophily, ego-network structures can help to recommend people to interact with [11, 17].

In our work, we propose *intermediary topics* as a feature to consider when recommending users to follow. The intuition behind intermediary topics is that they focus on homophily in specific shared interests that are non confronting nor challenging, *i. e.*, unlikely to provoke cognitive dissonance. Our definition of intermediary

*Corresponding author: eduardo.graells@telefonica.com. Work carried out while the first author was a PhD student at Universitat Pompeu Fabra, Barcelona, Spain.

topics is based on topic modeling using *Latent Dirichlet Allocation* [6, 23]. In particular, we build a *topic graph* of relations between latent topics, and find which ones are more likely to include people from diverse political backgrounds by estimating the information centrality [8] of latent topics.

Although topic modeling has been used before to measure homophily by considering user similarity [26], we measure its presence as the deviation from the expected interaction behavior given the population distribution in terms of user stances on specific controversial political issues. This distinction is important given that homophily also appears in other dimensions (*e. g.*, demography).

To study political leaning in social media, in particular in microblogging platforms, the first challenge is to actually detect what is the political leaning of users, as this attribute is not usually part of a public profile. One way to address the issue of classifying users is through supervised machine learning [13] and bayesian estimation [7], among other methods. Features used in classification include vocabulary, hashtags, and connectivity with accounts with known political leaning. Knowing political alignment of users allows to study group polarization. In a work related to our case study, Yardi *et al.* [27] studied debates about abortion in Twitter, in particular between users of *pro-life* and *pro-choice* stances. Their results indicate that the interaction between users having the same stance reinforced group identity, and discussions with members of the opposite group were found to be not meaningful, partly because the interface did not help in that aspect. In our work, we focus on a previously unexplored context: a politically centralized Latin-American country [15]. We complement previous work and help to understand the differences in political discussion around the globe.

3. METHODS

In this section we present our methodology to model users' intermediary topics, which extends previous work [16].

Sensitive Issues and Shared Interests. *Sensitive issues* are political or ideological topics for which their stances or opinions tend to divide people. This considers topics like *global warming*, *social security*, *health care reforms*, and *abortion*. Such topics tend to polarize people, *i. e.*, users who support one stance in abortion do not interact with users who support another stance, a behavior explained by homophily and cognitive dissonance. Conversely, *shared interests* are topics for which their stances or opinions do not, in normal conditions, tend to divide people. As example, people who support the soccer team *F.C. Barcelona* have a rivalry with people who support *Real Madrid F.C.*, however, the selective exposure mechanism would not be activated when discriminating information coming from people who support the opposite team–in fact, in some cases, they might be interested in such information. Other contexts can be less challenging as there might be no explicit rivalries. For instance, people with different musical tastes might be interested in discussing the particularities of their preferred music styles for comparison with others. As such, those shared interests could be good features to consider when introducing people [16], specially when considering first impressions [2].

Representation of User Stances in Sensitive Issues. An assumption we make with respect to user stances is that they are linked by partisan political ideology, *e. g.*, conservative/liberal people share views on different sensitive issues. Then, to estimate user stances, we first need to be able to estimate what users say with respect to sensitive issues. In Twitter, often users annotate their tweets with *hashtags*, which are text identifiers that start with the character #. For instance, *#prochoice* and *#prolife* are two hashtags related to two abortion stances, and each one of those stances has specific words related to them (*e. g.*, *"right to choose"* is pro-choice, and *"it is life since conception"* is pro-life). Pennacchiotti *et al.* [21] call those related words *prototypical words and hashtags*. We refer to both as prototypical keywords indistinctively. For any sensitive issue under consideration, we collect relevant tweets based on prototypical keywords (*e. g.*, *#prochoice*, *#prolife*, *abortion*, *pregnancy*, *interruption*, etc.). Those keywords can be extracted from a knowledge base of issues, with their respective related stances and associated terms. This knowledge base should be manually constructed to account for the social context of the population under study, as well as the contingency surrounding political discussion.

We build *user documents*, defined as the concatenation of tweets from each user. We represent each user document u as a vector

$$\vec{u} = [w_0, w_1, \ldots, w_n],$$

where w_i represents the vocabulary word i weighted using TF-IDF [3]:

$$w_i = freq(w_i, u) \times \log_2 \frac{|U|}{|u \in U : w_i \in u|},$$

where U is the set of users, and the vocabulary contains all prototypical keywords as well as all other words used by them. Note that the user document can be built with all tweets and retweets for each user, as well as a subset of both. In particular, we consider tweets and retweets, but not replies to other users, as they are less likely to contribute information to the document. Likewise, for each issue stance we build a stance vector $\vec{s}$, defined as the vectorized representation of tweets containing its prototypical keywords:

$$\vec{s} = [w_0, w_1, \ldots, w_n],$$

with w_i weighted according to TF-IDF with respect to the corpus of user documents.

Using these definitions we can estimate how similar is the language employed by a specific user with the known stances of a specific issue. Formally, we define a user stance with respect to a given sensitive issue as the feature vector $\vec{u}_s$ containing the similarity of user $\vec{u}$ with each issue stance. In this way, we consolidate all similarities in a *user stance vector*:

$$\vec{u}_s = [f_0, f_1, \ldots, f_{|S|}],$$

where S is the set of stances for the all sensitive issues under consideration, and f_i is the cosine similarity between $\vec{u}$ and the issue stance $\vec{s}_i$:

$$\text{cosine\_similarity}(\vec{u}, \vec{s}_i) = \frac{\vec{u} \cdot \vec{s}_i}{\| \vec{u} \| \| \vec{s}_i \|}.$$

Having this representation of user stances, we define the *view gap* with respect to a sensitive issue between two users as the distance between their respective user stance vectors.

Topic Graph. To build the topic graph, we rely on *Latent Dirichlet Allocation*. LDA is a generative topic model that clusters words based on their co-occurrences in documents, and defines latent topics that contribute words to documents. In the past, this model has given reliable results when applied to user documents. Thus, by using LDA we are able to estimate $P(t \mid u)$, for a given latent topic t and a given user document u from the set of users U. The topic graph is an undirected graph $G = \{T, V\}$, where the node set $T = \{t_0, t_1, \ldots, t_k\}$ is comprised of the k latent topics obtained from the application of LDA to the user documents, in the same way as Ramage *et al.* [23]. The edge set is defined as $V = \{v_{i,j} : P(t_i \mid u) \geq \epsilon \wedge \ P(t_j \mid u) \geq \epsilon \ \exists \ u \in U\}$, *i. e.*, two nodes are connected if both corresponding topics contribute (with a minimum probability

of ϵ) to the same user document. Note that edges are weighted according to the fraction of user documents that contributed to it.

Intermediary Topics. To estimate which topics are suitable to be used for recommendation of people of opposing views, we estimate the centrality of each node in the topic graph. In contrast to a previous definition of intermediary topics [16], instead of *betweenness centrality* we compute *current flow closeness centrality* [8] of nodes, which is equivalent to *information centrality* [25]. If the topic graph is considered as an electrical network, with edges replaced with resistances, information centrality is equivalent to the inverse of the average of correlation distances between all possible paths between two nodes. Using this analogy, we expect to measure the degree in which a topic might represent a shared non-challenging interest (*i. e.*, those with the least resistance) between two users. Hence, we redefine *intermediary topics* as topics whose centrality is higher than the median centrality of the entire graph.

In the next section, we evaluate this methodology through a case study of political discussion in Chile.

4. CASE STUDY: ABORTION IN CHILE

In this section we describe a case study where we analyze the issue of abortion in Chile using our methodology. In the context of ongoing campaigns for presidential elections, we crawled tweets from July 24th, 2013 to August 29th, 2013 using the *Twitter Streaming API*. Although we crawled tweets about general political discussion, we did focus crawling and analysis on *abortion*. After the analysis, we statistically evaluate intermediary topics to find how they differ in comparison to non-intermediary topics.

Why Abortion in Chile? The Duality in Discussion. The history of abortion in Chile is long, being declared legal in 1931 and illegal again in 1989. As of 2015, abortion is still illegal, making Chile one of the countries with most severe abortion laws in the world [24]. Abortion in Chile as a sensitive issue has good properties for analysis, as it is constantly being discussed in the political active population. On one hand, 61% of population was estimated to be catholic, and 21% professed another religion, while only 19% of the population was atheist or agnostic [22]. On the other hand, 63% of the Chilean population was in favor of legalization of abortion in 2013 [10]. The occurrence of several protests around public education, same-sex marriage and abortion, among other sensitive issues, are encouraging the usage of micro-blogging platforms and social networks to spread ideas and generate debates (for a discussion on the student movement in Chile see Barahona *et al.* [4]). There is a duality in how the country approach political issues. On one hand, a majority of the population is estimated to have conservative views. On the other hand, a majority of population is in favor of legalization of abortion. Because a growing portion of the population is asking for reforms using social media as a primary communication and organization device, Chile is an ideal scenario for our analysis.

4.1 Dataset Description

Query Keywords and Filtering. Initially, we used *query keywords* about known sensitive issues and hashtags: *abortion* (issue), *education* (issue), *same-sex marriage* (issue), *Sebastián Piñera* (president in 2013), *Michelle Bachelet* (candidate), *Evelyn Matthei* (candidate), among others. When crawling tweets we considered keywords about general political discussion and other sensitive issues (in addition to abortion) because we will consider the relationship between language usage and user stances. We also added emergent hashtags related to news events that happened during the crawling period. For instance, *#yoabortoel25* is about a protest held on July 25th [9]. Figure 1 shows the most frequent terms found in our collection.

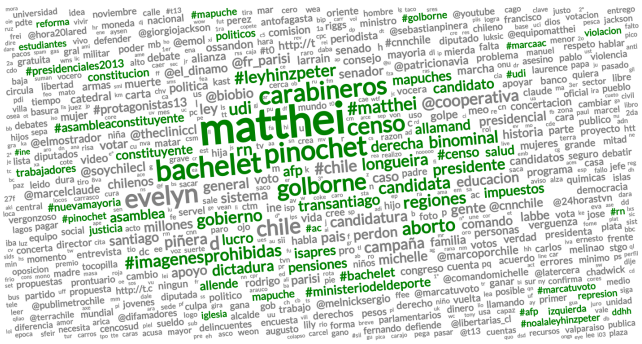

Figure 1: Wordcloud of frequent terms in the collection. Green terms were used as query keywords for crawling. Font size is proportional to frequency.

The most prominent words are last names of candidates, namely *Evelyn Matthei*, *Michelle Bachelet*, *Pablo Longueira* and *Laurence Golborne*. The last name of the dictator *Augusto Pinochet* is also prominent. Other prominent keywords are *carabineros* (the police), *censo* (the national level census conducted in 2012, with multiple flaws discovered in 2013), *Transantiago* (public transport system in Santiago), *isapres* (the private health system) and *AFP* (the name of the Chilean private pension system, composed of several *Administrators of Public Founds*). We filtered tweets in other languages than Spanish, tweets that were not geolocated to Chile according to users' self-reported location, as well as tweets about unrelated themes.

Dataset Size. In total, we analyzed 367,512 tweets about political discussion from 57,566 accounts that were geolocated in Chile using a gazetteer. Of those tweets, 18,148 are related to abortion, as they contain at least one prototypical keyword (see Table 1 for the list of keywords related to abortion). The vocabulary size is 38,827, filtering out all keywords that appear in less than 5 tweets.

Pro-Choice and Pro-Life Stances. We manually built a list of words, accounts, and hashtags related to abortion and its two stances. We iteratively explored the dataset to find co-occurrences of prototypical keywords like *abortion*, *#abortolibre* (*free abortion*) and *#noalaborto* (*no to abortion*). For pro-choice and pro-life keywords, the number of seed users and their number of tweets are displayed. These seeds represent whether a user document contained keywords from one stance but not from the other, *e. g.*, a user document that contains at least one pro-choice keyword and no pro-life keywords is considered a pro-choice seed user. As observed in Table 1, the number of pro-choice seed users outnumbers those of pro-life stance (1,934 pro-choice against 338 pro-life). This does not necessarily indicate the proportion of users from both stances. For instance, after performing a manual exploration, some pro-life users who identify themselves as pro-life in their biographies, tend to inject content into pro-choice timelines by publishing tweets with prototypical hashtags from the opposite stance [12].

To build the stance vectors of pro-choice and pro-life stances, we concatenated the tweets of the corresponding seed users of each stance. Then, according to our methodology, we estimated the user stances on abortion by computing the cosine similarity between user vectors and the stance vectors. These similarities are displayed with hexagonal binning in Figure 2, where the x axis represents similarity with the pro-choice stance vector $\vec{s_c}$; and the y axis represents similarity with the pro-life stance vector $\vec{s_\ell}$. We display two charts: one for users who have tweeted about abortion (8,794) on the left, and one that considers all users on the dataset (57,566) on the right. This

Table 1: Keywords used to characterize the pro-choice and pro-life stances on abortion. General keywords plus stance keywords were used to find people who talked about abortion in Twitter. Seeds are users who published tweets with keywords from only one abortion stance.

Stance	Tweets	Seeds	Keywords
Pro-choice	95,173	1,934	#abortolibre, #yoabortoel25, #abortolegal, #yoaborto, #abortoterapeutico, #proaborto, #abortolibreseguro gratuito, #despenalizaciondelaborto, #abortoetico, #abortolegal, #abortosinapellido, #derechoadecidir
Pro-life	10,040	338	#provida, #profamilia, #abortoesviolencia, #noalaborto, #prolife, #sialavida, #dejalolatir, #siempreporlavida, provida, #nuncaaceptaremoselaborto, #chilenoquiereabortos, #conabortonohayvoto, #yoasesinoel25, #somosprovida
General Words	–	–	aborto(s), abortista(s), abortados(as), abortivo(a)... ...(tenses of *to abort* in spanish)
Related Hashtags	–	–	#marchaabortolegal, #bonoaborto, #cifrasaborto, #feminismo
Relevant Accounts	–	–	@elardkoch, @siemprexlavida, @quieronacer, @mileschile, @melisainstitute, @ObservatorioGE
Contingency Words	–	–	terapéutico, violada, violación, violaciones, interrupción, inviabilidad, embarazo, embarazada, feto, embrión, fecundación, antiaborto, feminismo

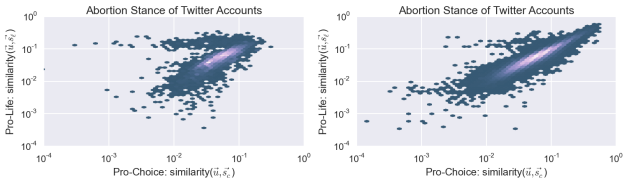

Figure 2: Distributions of user stances based on similarity between user vectors and stance vectors (pro-life and pro-choice). Left: stances of users who tweeted about abortion. Right: stances of all users in the dataset. Both charts use a log-log scale.

is possible because the user stance vectors are constructed using all the vocabulary employed by seed users; hence, they contain valid weights for words unrelated to abortion, but related to additional issues that those users discussed. Under the assumption that sensitive issues have a degree of correlation among stances in different issues, this allows us to estimate a tendency for all users. We define *stance tendency* as:

$$\text{tendency} = \text{cosine\_similarity}(\vec{u}, \vec{s_c}) - \text{cosine\_similarity}(\vec{u}, \vec{s_\ell}).$$

We classify users with tendency ≥ 0 as pro-choice, and pro-life otherwise. The median stance tendency is 0.02, showing a slight tendency towards the pro-choice stance: 54.98% of users are classified as pro-choice, while 45.02% of users are classified as pro-life. Pro-choice users published 10.24 tweets in average, while pro-life users published 10.48 tweets in average.

According to the *Center of Public Studies* [10], 63% of the Chilean population was in favor of legalization of abortion in 2013. Our predicted proportion of user stances does not differ from expectations according to a chi-square test ($\chi^2 = 2.76$, p $= 0.10$). While the Twitter population is not demographically representative of the population, this result indicates that abortion stances are reflected on the micro-blogging platform Twitter.

4.2 Homophily in Two-Way Interactions

Having predicted a stance for each user in the dataset, we are able to evaluate if the interactions in the dataset are homophilic, *i. e.*, we test if users tend to interact with people of the same abortion stance. To do so, we study 2-way interactions. Mentions and retweets are 1-way interactions, where the target user is not necessarily a participant of the interaction. When the target user replies to the mention or the retweet, we consider it a 2-way, bidirectional interaction. To measure homophily, we estimate the aggregated interactions between users in both stances, and compare their inter-stance proportions with the proportions of predicted stances for all

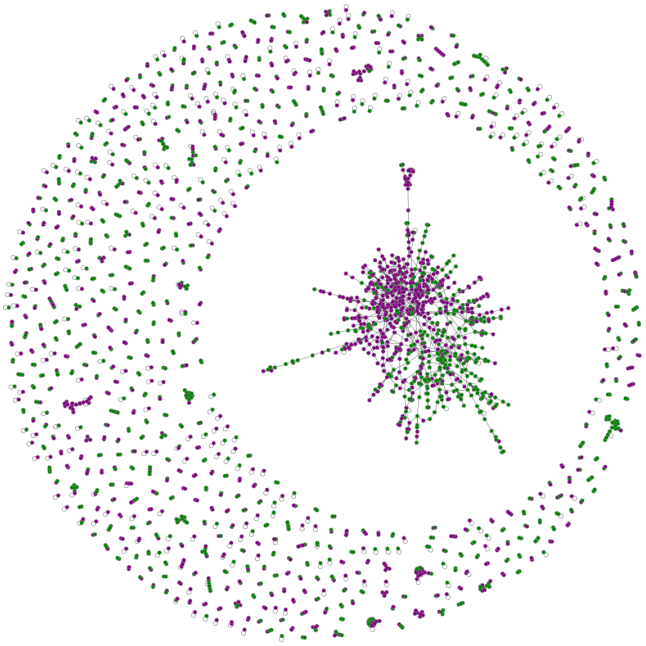

Figure 3: A spring-based graph visualization of two-way user interactions in abortion discussion, where nodes are users. Color encodes abortion stance (purple: pro-choice; green: pro-life).

accounts. If the interaction behavior is unbiased, then the proportion of interactions between stances should not differ from the proportion of users in each stance.

To avoid bias in the estimation, we only considered each pair (u_1, u_2) once per inter-stance interactions. The number of 2-way interactions found for each stance is: pro-choice, 2,234; pro-life, 2,042. The structure of those interactions is visualized in Figure 3, where it can be observed that the largest component has two identifiable clusters, and that small components are prominently of one stance only. The proportions of interactions with the same stance is similar (pro-choice: 76.45%; pro-life: 74.24%). Given the distribution of user stances, in an unbiased population we would expect that each stance would have bidirectional interactions distributed according to the population, *e. g.*, 54.98% of pro-choice users' interactions would be with those of the same stance. A chi-square test indicates that both proportions differ significantly from the expectations (pro-life: $\chi^2 = 29.55$, p < 0.001, Cohen's $w = 0.33$; pro-choice: $\chi^2 = 22.91$, p < 0.001, Cohen's $w = 0.31$), confirming homophilic behavior in the studied population.

4.3 Intermediary Topics

Of all Chileans who published tweets in the case study, we selected a group of 4,077 candidates for analysis of intermediary topics. We considered users that were likely to be *regular users*, *i. e.*, those who follow less than 2,000 accounts and are followed by less than 2,000 (a limit defined by Twitter). This filtering was made because regular people are arguably more prone to discuss their own interests, unlike popular accounts which may be from media outlets, blogs, or celebrities. From those users, we crawled 1,400,582 tweets from December 6th, 2013 until January 3rd, 2014. Jointly with our abortion stance estimation of those users, this makes this dataset useful to test the political diversity of intermediary topics.

We ran LDA with $k = 200$ (a value used before in similar contexts [23]), built the topic graph and estimated information centrality as defined by our methodology. After removing junk topics, which do not contribute to any user document, the graph contains 198 nodes and 6,906 edges. The median centrality is 1.23×10^{-4}, and its maximum value is 1.64×10^{-4}.

We analyze three variables and their relation with centrality, as well as their differences between intermediary and non-intermediary topics: the percent of users that each topic contributes to (Figure 4 Left); the probability of abortion keywords to contribute to each topic (Figure 4 Right), estimated using the LDA model; and the stance diversity (Figure 4 Center), which is the *Shannon entropy* [18] with respect to the predicted abortion stances for all users related to a topic:

$$\text{diversity} = \frac{-\sum_{i=1}^{|S|} p_i \ \ln p_i}{\ln |S|},$$

where S is the set of stances, and p_i is the probability of stance i, estimated from the fraction of users assigned to each stance according to our methodology.

Proportion of Users. Central topics have much more users than non-central ones: as the number of users increment, centrality does. This is confirmed by a Spearman ρ rank-correlation of 0.99 ($p < 0.001$) between proportion of users and centrality. The maximum proportion of users a topic contributes to is 78.78%, the median value is 0.56% and the mean is 4.13%. The mean for intermediary topics is 7.99%, and for non-intermediary topics 0.26%. This difference is significant according to a Mann-Whitney U test ($U = 12.10$, $p < 0.001$). Hence, intermediary topics are more populated than non-intermediary topics. This is an expected result, because topic graph construction is based on how topics are related to users.

Stance Diversity. Nodes with high stance diversity can have low centrality, but they concentrate in the upper middle of the chart. The maximum diversity of a topic is 1, its median value is 0.97 and its mean is 0.91. The mean for intermediary topics is 0.96, and for non-intermediary topics 0.86. This difference is significant according to a Mann-Whitney U test ($U = 3.30$, $p < 0.001$), meaning that intermediary topics are more likely to contain a greater diversity of people with different views on abortion than non-intermediary topics.

Topical Probability of Abortion-Related Vocabulary. Using our set of prototypical keywords, we can estimate the probability of abortion-related vocabulary to contribute to specific topics $P(A \mid t)$, where A is the set of keywords, and t is the target topic:

$$P(A \mid t) = \sum_{i=1}^{|A|} P(w_i \mid t),$$

where w_i is the ith word in A. Note that the LDA model allows us to estimate $P(w_i \mid t)$ directly. Figure 5 displays the distributions and

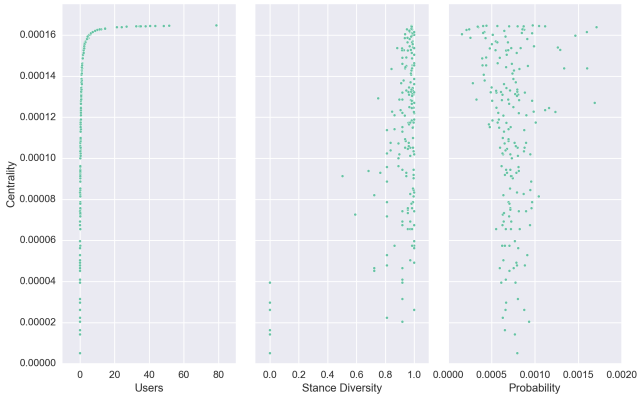

Figure 4: Relationship between topic information centrality [8] and the percent of users the topic contributes to (left), the abortion-stance diversity estimated with *Shannon entropy* [18] (center), and the probability of abortion-related keywords to contribute to each topic (right).

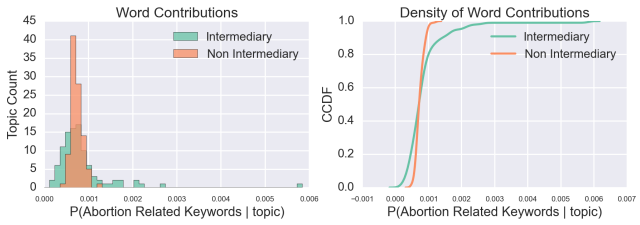

Figure 5: Left: Histograms of abortion-related keywords contributions to intermediary and non-intermediary topics. Right: Cumulative Density Function .

Complimentary Cumulative Density Functions (CCDFs) of probabilities for intermediary and non-intermediary topics. Although the distribution chart hints a potential difference, this difference is not significant according to a Mann-Whitney U test ($U = -0.59$, $p = 0.55$).

5. DISCUSSION

In this paper we have confirmed that intermediary topics do exist and are measurable. We have improved the definition of intermediary topics by Graells-Garrido *et al.* [16], as well as quantified homophilic discussion and the differences between intermediary and non-intermediary topics. In particular, we have found that intermediary topics are more likely to contain a diverse set of users in terms of political stances, and thus, are suitable for use in recommendation of people of opposing views. We devise these topics as important features that could help to avoid *cognitive dissonance* [14] in users when facing recommendations. Although our results apply to the studied community from Chile, the methods used are generalizable to other communities as long as there are known prototypical keywords for the sensitive issues to be studied.

In addition, the way in which we quantified homophily can be used as a metric to evaluate the polarization in discussion around specific political issues. In our case study, polarization of stances had considerable effect sizes (measured with Cohen's w), meaning that discussion in Chile around abortion is highly polarized, a result supported by national surveys of political discussion [22, 10].

A question that arises regarding intermediary topics is: does the definition of intermediary topics hold when considering general

political views instead of a specific sensitive issue? We propose that it does because by definition intermediary topics only rely on the estimation of information centrality [8]. However, this is left for future work. Additionally, future work will consider the incorporation of intermediary topics into a recommender system to be evaluated with users, as well as the interaction of intermediary topics with social- and content-based signals.

Acknowledgments. We thank the anonymous reviewers for their helpful feedback. This work was partially funded by Grant TIN2012--38741 (Understanding Social Media: An Integrated Data Mining Approach) of the Ministry of Economy and Competitiveness of Spain.

References

[1] Jisun An, Daniele Quercia, and Jon Crowcroft. "Why individuals seek diverse opinions (or why they don't)". In: *Proceedings of ACM Web Science*. 2013, pp. 11–15.

[2] Solomon E Asch. "Forming impressions of personality." In: *The Journal of Abnormal and Social Psychology* 41.3 (1946), p. 258.

[3] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information retrieval: the concepts and technology behind search, 2nd. Edition*. Addison-Wesley, Pearson, 2011.

[4] Matías Barahona, Cristóbal García, Peter Gloor, and Pedro Parraguez Ruiz. "Tracking the 2011 student-led movement in Chile through social media use". In: *Collective Intelligence 2012* (2012).

[5] Pablo Barberá. "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data". In: *Political Analysis* 23.1 (2015), pp. 76–91.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *The Journal of Machine Learning Research* 3 (2003), pp. 993–1022.

[7] Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. "What's in Twitter, I know what parties are popular and who you are supporting now!" In: *Social Network Analysis and Mining* 3.4 (2013), pp. 1379–1391.

[8] Ulrik Brandes and Daniel Fleischer. "Centrality measures based on current flow". In: *STACS 2005* (2005), pp. 533–544.

[9] María Jesús Ibáñez Canelo. "El control de los cuerpos de las mujeres es algo medular en la política patriarcal capitalista: entrevista a Soledad Rojas, feminista chilena". In: *Comunicación y Medios* 30 (2015). [In Spanish. Title translation: The control of women's bodies is something core in capitalist patriarchal politics: interview with Soledad Rojas, Chilean feminist.]

[10] CEP. *National Survey of Public Opinion, September–Octuber 2013*. http://www.cepchile.cl/1_5388/doc/estudio_nacional_de_opinion_publica_septiembre-octubre_2013.html. [Online; accessed April 2015]. 2013.

[11] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. "Make new friends, but keep the old: recommending people on social networking sites". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2009, pp. 201–210.

[12] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. "Political polarization on twitter." In: *International Conference on Weblogs and Social Media*. 2011.

[13] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. "Predicting the political alignment of twitter users". In: *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*. IEEE. 2011, pp. 192–199.

[14] Leon Festinger. *A theory of Cognitive Dissonance*. Vol. 2. Stanford University Press, 1962.

[15] Eduardo Graells-Garrido and Mounia Lalmas. "Balancing Diversity to Counter-measure Geographical Centralization in Microblogging Platforms *(short paper)*". In: *25th ACM Conference on Hypertext and Social Media* (2014).

[16] Eduardo Graells-Garrido, Mounia Lalmas, and Daniele Quercia. "People of opposing views can share common interests". In: *Proceedings of the companion publication of the 23rd international conference on World Wide Web (poster)*. International World Wide Web Conferences Steering Committee. 2014, pp. 281–282.

[17] John Hannon, Mike Bennett, and Barry Smyth. "Recommending twitter users to follow using content and collaborative filtering approaches". In: *Proceedings of the fourth ACM Conference on Recommender Systems*. ACM. 2010, pp. 199–206.

[18] Lou Jost. "Entropy and diversity". In: *Oikos* 113.2 (2006), pp. 363–375.

[19] Miller McPherson, Lynn Smith-Lovin, and James M Cook. "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* (2001), pp. 415–444.

[20] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[21] Marco Pennacchiotti and Ana-Maria Popescu. "Democrats, republicans and starbucks afficionados: user classification in twitter". In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. ACM. 2011, pp. 430–438.

[22] Pontificia Universidad Católica de Chile. *Encuesta Bicentenario UC-Adimark, 2014*. http://encuestabicentenario.uc.cl/. [In Spanish; Online; accessed April 2015]. 2014.

[23] Daniel Ramage, Susan Dumais, and Dan Liebling. "Characterizing microblogs with topic models". In: *International Conference on Weblogs and Social Media*. Vol. 5. 4. 2010, pp. 130–137.

[24] Bonnie L Shepard and Lidia Casas Becerra. "Abortion policies and practices in Chile: ambiguities and dilemmas". In: *Reproductive Health Matters* 15.30 (2007), pp. 202–210.

[25] Karen Stephenson and Marvin Zelen. "Rethinking centrality: Methods and examples". In: *Social Networks* 11.1 (1989), pp. 1–37.

[26] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. "Twitterrank: finding topic-sensitive influential twitterers". In: *Proceedings of the third ACM WSDM*. 2010, pp. 261–270.

[27] Sarita Yardi and Danah Boyd. "Dynamic debates: An analysis of group polarization over time on twitter". In: *Bulletin of Science, Technology & Society* 30.5 (2010), pp. 316–327.